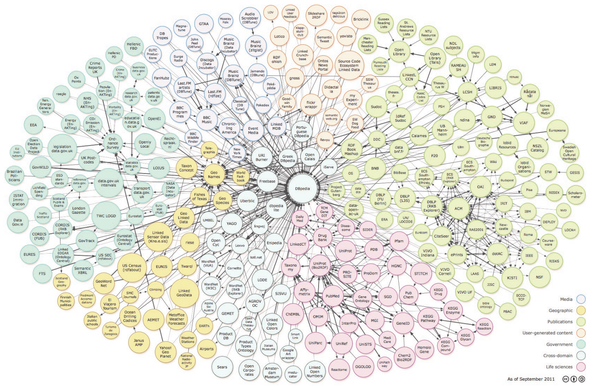
Mining Linked Open Data: a Case Study with Genes Responsible for Intellectual Disability

Gabin Personeni, Simon Daget, Céline Bonnet, Philippe Jonveaux, Marie-Dominique Devignes, Malika Smaïl-Tabbone, Adrien Coulet

Mining Linked Open Data: a Case Study with Genes Responsible for Intellectual Disability

Gabin Personeni[1], Simon Daget[1], Céline Bonnet[2], Philippe Jonveaux[2],
Marie-Dominique Devignes[1], Malika Smaïl-Tabbone[1], Adrien Coulet[1]
1. LORIA (CNRS, Inria NGE, Université de Lorraine, FRANCE) 2. Laboratoire de Génétique Médicale, CHU de Nancy, FRANCE, INSERM U-954
Contact : gabin.personeni@loria.fr



Motivation

Increasing amounts of life sciences data are made available through the **Linked Open Data** (Bio2RDF, EMBL-EBI RDF platforms).
Challenge: use Linked Open Data to answer a biological question.
Approach:

1. query and integrate Linked Open Data,
2. mine the collected data using **Inductive Logic Programming**, a relational data mining method.
3. assess the contribution of domain knowledge available in the Linked Open Data to the quality of both characterization and prediction.

Application: characterize **genes responsible for Intellectual Disability** and build a model predicting whether a gene is responsible for Intellectual Disability or not.

Biological Question

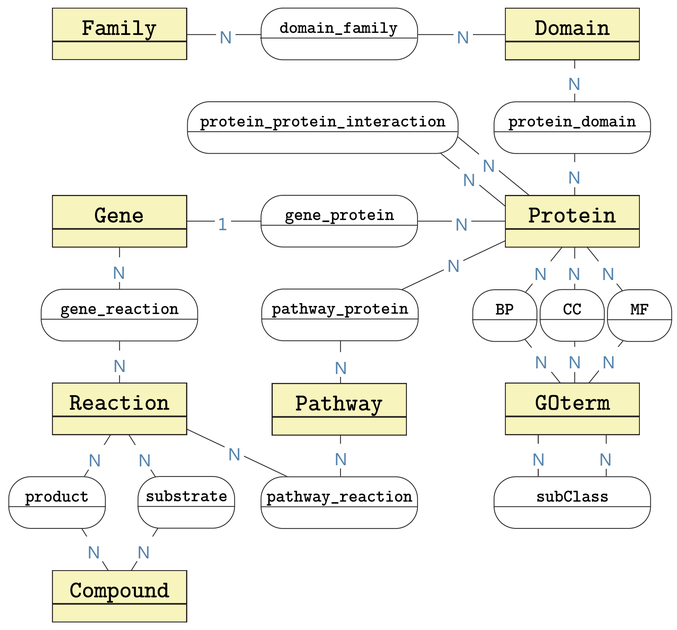
Concept to learn

Characterization of genes responsible for Intellectual Disability (ID).
Input :
-list of 282 genes responsible for ID (Inlow and Restifo, 2004).
-list of 267 genes responsible for other phenotypes according to OMIM.
-data on these genes available in the Linked Open Data (LOD).

Formalization

Data Model

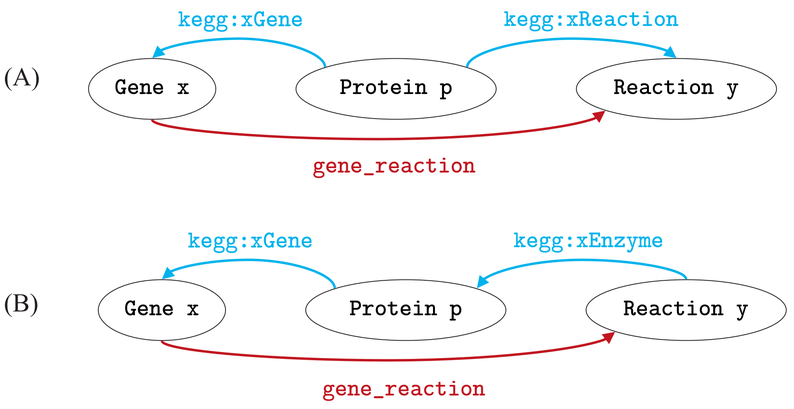
Entities and relationships



Entity-Relationship Model covering the data related to the biological question

-Restricted to binary relationships to fit the LOD
-Domain knowledge in the form of the Gene Ontology *is-a* hierarchy

Mappings to the Linked Open Data



Mapping of the `gene_reaction` relationship.

-Blue: RDF properties. Red: E/R relationship definition.

A gene `x` is related to a reaction `y` if there is a protein `p` whose coding gene (`kegg:xGene` property) is `x` and either:

(A) `p` is involved in `y` (`kegg:xReaction` property),
(B) `y` has `p` for an enzyme (`kegg:xEnzyme` property)

SPARQL Queries

```
PREFIX kegg:<http://bio2rdf.org/kegg_vocabulary:>
PREFIX geneid:<http://bio2rdf.org/geneid_vocabulary:>
SELECT ?x ?y
WHERE
{
   {?x rdf:type geneid:Gene}
   UNION                                  } Mapping of Gene
   {?x rdf:type kegg:Gene}

   {?y rdf:type kegg:Reaction}            } Mapping of Reaction

   ?p kegg:xGene ?x.
   {?y kegg:xEnzyme ?p}                   } Mapping of gene_reaction
   UNION
   {?p kegg:xReaction ?y}
}
```

One query per relationship: example with the gene_reaction relationship

-Query for relationship R retrieves couples (x,y) such that $R(x,y)$
-Involves the mapping of the relationship, plus the mapping of its domain and range
-Apply restrictions, for instance, domain restriction on our list of genes

Data Integration

Triple Store

Relationship	Triples	Relationship	Triples	Entity	Instances
subClass	12779	pathway_protein	767	GOterm	7770
protein_BP	10242	pp_interaction	742	Protein	1257
protein_CC	4358	gene_reaction	500	Family	781
protein_MF	4063	protein_domain	262	Compound	628
domain_family	1238			Pathway	580
product	960			Gene	549
substrate	938			Reaction	433
gene_protein	819			Domain	262

Amounts of collected data per relationship and entity

Data Mining

Relational Data Mining with Inductive Logic Programming (ILP) using the Aleph program

Principles: generalization of a positive example (gene responsible for ID) to build a rule covering the most positive examples, and the least negative examples

Advantages: ILP natively supports the relational format of the LOD data, and can perform inference on domain knowledge (GO *is-a* hierarchy)

Experiments:

-*no-GO*: all data except GO annotation data

-*GO1*: all data including GO annotations, no inference on the GO *is-a* hierarchy (only one level is used)

-*GO2*, *GO3* and *GO4*: all data including GO annotations, inferences over 2, 3, and 4 levels of the GO *is-a* hierarchy, respectively

Theory

First-order logic rules

Characterization:

-*no-GO* theory is highly specific: points to individual reactions and pathways, low amount of false positives, but low coverage of positive examples

-*GO1-4* theories rarely use other data than GO annotations, but better coverage of positive examples

Prediction evaluation:

-Scores computed using leave-one-out cross-validation (KNIME workflows)
-Sensitivity and accuracy improves with each additional inference level

Rule	Covered positive, negative examples	
`is_responsible(A):-gene_in_reaction(A,B), gene_ch(A,x).`	15	2
`is_responsible(A):-gene_in_reaction(A,B), gene_ch(A,'1').`	14	0
`is_responsible(A):-gene_in_pathway(A,'Valine, leucine and isoleucine degradation').`	11	1
`is_responsible(A):-gene_in_pathway(A,'N-Glycan biosynthesis').`	8	0
`is_responsible(A):-gene_in_pathway(A,'Glycosaminoglycan degradation').`	8	0
`is_responsible(A):-gene_in_reaction(A,'Ubiquinol + Acceptor <=> Ubiquinone + Reduced acceptor').`	7	0
`is_responsible(A):-gene_in_reaction(A,B), gene_protein(A,C), pp_interaction(C,P30480).`	7	0
`is_responsible(A):-gene_chromosome_band(A,'22q13').`	6	0
`is_responsible(A):-gene_in_reaction(A,B), gene_protein(A,C), pp_interaction(C,D), pp_interaction(D,C).`	6	0
`is_responsible(A):-gene_in_pathway(A,'Alanine and aspartate metabolism').`	6	1
`is_responsible(A):-gene_in_pathway(A,'Formation of transcription-coupled NER (TC-NER) repair complex').`	5	0

***no-GO* theory**

Rule	Covered positive, negative examples	
`is_responsible(A):-gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,D), subClass(D,'organonitrogen compound catabolic process').`	42	2
`is_responsible(A):-gene_protein(A,B), protein_bp(B,C), subClass(C,'cellular amino acid metabolic process'), subClass(C,'cellular metabolic process').`	32	3
`is_responsible(A):-gene_in_reaction(A,B), gene_protein(A,C), protein_bp(C,'small molecule metabolic process'), protein_cc(C,'mitochondrial inner membrane').`	23	1

Best 3 rules out of 16 from the *GO4* theory

Experiment	Sensitivity (%) TP/P	Specificity (%) TN/N	Accuracy (%) (TP+TN)/(P+N)
no-GO	26.6	94.4	59.6
GO1	47.9	81.3	64.1
GO2	55.7	80.5	67.8
GO3	55.7	81.7	68.3
GO4	57.1	83.1	69.8

Leave-one-out cross-validation evaluation for each experiment

References

-Gabin Personeni *et al*. Mining Linked Open Data: a Case Study with Genes Responsible for Intellectual Disability. In DILS. Springer, 2014.

-Jennifer K Inlow and Linda L Restifo. Molecular and comparative genetics of mental retardation. Genetics, 166 (2):835—881, 2004.

-Ashwin Srinivasan. The Aleph Manual, 2007: http://www.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/

-KNIME website: http://www.knime.org

-Renaud Grisoni *et al*. Méthodologie et outils pour l'extraction de connaissances par Programmation Logique Inductive (PLI). 13ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances, Toulouse, 29 janvier—1er février 2013 (poster).